Multiple Regression Analysis : Further Issues

1 Data scaling on OLS statistics

When we change the unit of measurment of a variable, the value of estimators would chance accordingly. For example

$$\widehat{bweght}_g = \widehat{\beta}_0 + \widehat{\beta}_1 cigs + \widehat{\beta}_2 fa\min c,$$

where
$bwght$ = child birth weight, in grams.
$cigs$ = number of cigarettes smoked by the mother while pregnant, per day.
$fa\min c$ = annual family income, in thousands of dollars.

- What if we use bweght in kilograms?

$$1 \text{ kg.} = 1000 \text{ g.}$$

$$\widehat{bweght}_{kg} = \frac{\widehat{bweght}_g}{1000} = \frac{\hat{\beta}_0}{1000} + \frac{\hat{\beta}_1}{1000} cigs + \frac{\hat{\beta}_2}{1000} faminc$$

$$= \hat{\alpha}_0 + \hat{\alpha}_1 cigs + \hat{\alpha}_2 faminc$$

$$\Rightarrow \hat{\alpha}_0 = \frac{\hat{\beta}_0}{1000}, \quad \hat{\alpha}_1 = \frac{\hat{\beta}_1}{1000}, \quad \hat{\alpha}_2 = \frac{\hat{\beta}_2}{1000}$$

- What if we use faminc in USD (instead of 1000 USD)

$$bweght_g = \hat{\beta}_0 + \hat{\beta}_1 cigs + \frac{\hat{\beta}_2}{1000} faminc_{USD}$$

The value of this variable is going to be 1000 times larger than faminc

$$= \hat{\beta}_0 + \hat{\beta}_1 cigs + \hat{\theta}_2 faminc_{USD}$$

$$\Rightarrow \hat{\theta}_2 = \frac{\hat{\beta}_2}{1000}$$

in other words $\hat{\theta}_2$ = impact of 1 USD ↑ in income
$\hat{\beta}_2$ = —— " —— 1000 USD ↑ in income

- What if we use bweght in kg & income in THB

$$bweght_{kg} = \frac{\hat{\beta}_0}{1000} + \frac{\hat{\beta}_1}{1000} cigs + \frac{\left(\frac{\hat{\beta}_2}{1000}\right)}{30{,}000} faminc$$

THB

This value is going to be 30,000 times more than faminc

2 More on functional forms

$\Delta y = y_1 - y_2$

$\Delta x = x_{11} - x_{12}$

! Logarithmic Functional Form

usually means natural log

$$\log(y) = \beta_0 + \beta_1 \log(x_1) + \beta_2 x_2 + u$$

$$\circ\ \beta_1 = \frac{d\log(y)}{d\log(x_1)} = \frac{\frac{1}{y}dy}{\frac{1}{x_1}dx_1} = \frac{\frac{1}{y}\Delta y}{\frac{1}{x_1}\Delta x_1} = \frac{100 \times \frac{1}{y}\Delta y}{100 \times \frac{1}{x_1}\Delta x_1} = \frac{\%\Delta y}{\%\Delta x}$$

with the log y & log x format, the coefficient is going to be the elasticity! (x_1 elasticity of y)

$$\circ\ \beta_2 = \frac{d\log(y)}{dx_2} = \frac{\frac{1}{y}dy}{dx_2} = \frac{\frac{1}{y}\Delta y}{\Delta x_2}$$

⇒ if we want the upper term to be % change, then

$$100\beta_2 = \frac{100\frac{1}{y}\Delta y}{\Delta x_2}$$

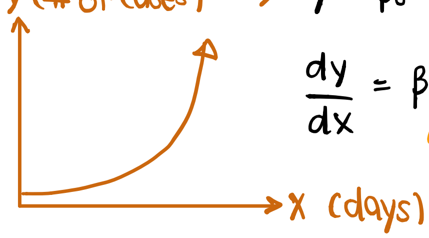
$$100\beta_2 = \frac{\%\Delta y}{\Delta x_2}$$

$100\beta_2 = \%\Delta$ in y given that x_2 increases by 1 unit.

! Models with Quadratics (squares)

→ Capture increasing / decreasing marginal effects (slope of the relationship between x & y is not constant.)

COVID-19 example

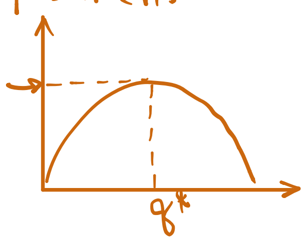


$$\Rightarrow y = \beta_0 + \beta_1 x + \beta_2 x^2 + u$$

$$\frac{dy}{dx} = \beta_1 + 2\beta_2 x$$

(+) (+) days

Decreasing returns



$$\Rightarrow y = \beta_0 + \beta_1 x + \beta_2 x^2 + u$$

$$\frac{dy}{dx} = \beta_1 + 2\beta_2 x$$

(+) (−)

$\pi = (P - MC)q$; $MC = 10$

Demand: $P = 100 - q$

$\pi = (100 - q - 10)q$

β_1 is positive

F.O.C $\frac{\partial \pi}{\partial q} = 0 = 90 - 2q$

Example : Effects of Pullution on Housing Prices

$$\log(price) = \beta_0 + \underline{\beta_1} \log(nox) + \underline{\beta_2} \log(dist) + \underline{\beta_3} rooms + \underline{\beta_4} room^2 + \underline{\beta_5} stratio + u$$

β_2 is (−)

Derivatives of exponential and logarithmic functions [edit]

$$\frac{d}{dx}\left(c^{ax}\right) = ac^{ax}\ln c, \qquad c > 0$$

the equation above is true for all c, but the derivative for $c < 0$ yields a complex number.

$$\frac{d}{dx}\left(e^{ax}\right) = ae^{ax}$$

$$\frac{d}{dx}\left(\log_c x\right) = \frac{1}{x\ln c}, \qquad c > 0, c \neq 1$$

the equation above is also true for all c, but yields a complex number if $c < 0$

$$\frac{d}{dx}\left(\ln x\right) = \frac{1}{x}, \qquad x > 0.$$

$$\frac{d\ln x}{dx} = \frac{1}{x} \Rightarrow \boxed{d\ln(x) = \frac{1}{x}dx}$$

$$\frac{d}{dx}\left(\ln |x|\right) = \frac{1}{x}.$$

$$\frac{d}{dx}\left(x^x\right) = x^x(1+\ln x).$$

$$\frac{d}{dx}\left(f(x)^{g(x)}\right) = g(x)f(x)^{g(x)-1}\frac{df}{dx} + f(x)^{g(x)}\ln\left(f(x)\right)\frac{dg}{dx}, \qquad \text{if } f(x) > 0, \text{ and if } \frac{df}{dx} \text{ and } \frac{dg}{dx} \text{ exist.}$$

$$\frac{d}{dx}\left(f_1(x)^{f_2(x)^{(\ldots)^{f_n(x)}}}\right) = \left[\sum_{k=1}^{n}\frac{\partial}{\partial x_k}\left(f_1(x_1)^{f_2(x_2)^{(\ldots)^{f_n(x_n)}}}\right)\right]\Bigg|_{x_1=x_2=\ldots=x_n=x}, \text{ if } f_{i<n}(x) > 0 \text{ and } \frac{df_i}{dx} \text{ exists.}$$

Logarithmic derivatives [edit]

The logarithmic derivative is another way of stating the rule for differentiating the logarithm of a function (using the chain rule):

$$(\ln f)' = \frac{f'}{f} \qquad \text{wherever } f \text{ is positive.}$$

Logarithmic differentiation is a technique which uses logarithms and its differentiation rules to simplify certain expressions before

where
$price$ = housing price
nox = level of pollution
$dist$ = distance from downtown
$rooms$ = number of rooms
$stratio$ = average student per teacher ratio
The estimation result is given by

In the US or many other countries, students can apply to schools in the area without having to take any test.
So, the lower stratio, the better the school.

```
regress  lprice lnox dist rooms rooms_sq stratio
```

Source	SS	df	MS
Model	51.4933152	5	10.298663
Residual	33.0889098	500	.06617782
Total	84.582225	505	.167489554

Number of obs =	506
F(5, 500) =	155.62
Prob > F =	0.0000
R-squared =	0.6088
Adj R-squared =	0.6049
Root MSE =	.25725

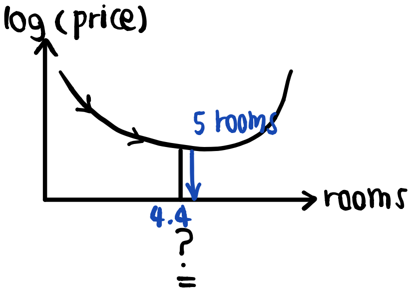
log(price) → lprice
log(nox) → lnox

lprice	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lnox	β_1 -.9767545	.0995938	-9.81	0.000	-1.172429	-.7810806
dist	β_2 -.0321972	.0094013	-3.42	0.001	-.050668	-.0137264
rooms	β_3 -.5528032	.1612965	-3.43	0.001	-.8697056	-.2359007
rooms_sq	β_4 .0624697	.0124867	5.00	0.000	.0379368	.0870025
stratio	β_5 -.0486667	.0058131	-8.37	0.000	-.0600879	-.0372455
_cons	13.59154	.5650901	24.05	0.000	12.4813	14.70178

$|t| > 1.96$ ↑ ↑ all < 0.05
⇝ all variables are significant

Consider the effect of "room"

$$\frac{d\log(price)}{d\,rooms} = \beta_3 + 2\beta_4 rooms = -0.553 + 2(0.062)\cdot rooms$$



* at how many rooms does 1 additional room has a positive impact on log(price)?

$$0 = -0.553 + 2(0.062)\cdot rooms$$
$$rooms = 4.4$$

Answer ⇒ at 4.4 rooms or more
at ⇝ 5 rooms or more

What wold be the % change in price when the number of room increases from 5 to 6?

- $\dfrac{d\log(price)}{d\,rooms} = -0.553 + 2(0.062)\cdot rooms$

$$\frac{100 \cdot \frac{1}{price}\, dprice}{d\,rooms} = 100(-0.553 + 2(0.062)\cdot 5)$$
$$= 100 \times 0.067 = 6.7\% \text{ increase}$$

total %Δ in price when # rooms ↑ from 5 to 7 is 6.7 + 19.1% = 25.8%

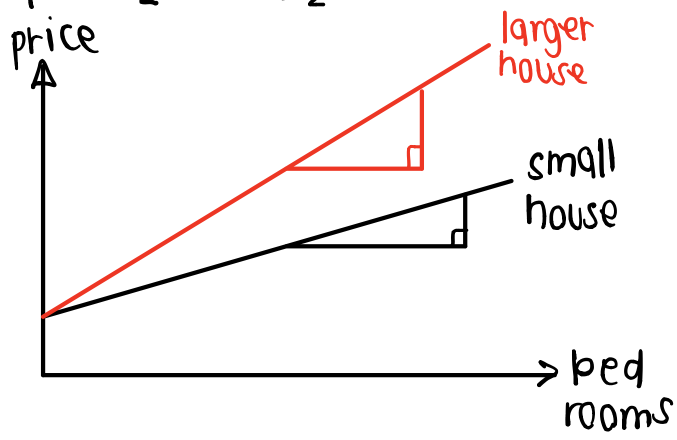
⇒ what about % in price when # rooms increases from 5 to 7?

$$\%\Delta price = 100(-0.553 + 2(0.062)\cdot 6) = 19.1\%$$

3 Models with Interaction Terms ⇒ used when the impact of one variable depends on the value (level) of another variable.

Consider

$$price = \beta_0 + \beta_1 \underset{X_1}{sqrft} + \beta_2 \underset{X_2}{bdrms} + \beta_3 \overbrace{\underset{X_1}{sqrft} \cdot \underset{X_2}{bdrms}}^{X_3} + \beta_4 \underset{X_2}{bthrms} + u$$



where
  $price$ = housing price
  $sqrft$ = house size (square feet)
  $bdrms$ = number of bedrooms
  $bthrms$ = number of bathrooms

- $\dfrac{d\,price}{d\,bdrms} = \beta_2 + \beta_3\, sqrft$

⇒ if $\beta_2 > 0$ then, an additional bedroom would increase price <u>more</u> for a larger house!

4 More on the Goodness-of-Fit and Selection of Regressors

! Adding more regressors ALWAYS improve fit ⤳ R^2 always ↑

- But we lose the "degree of freedom"
  (d.f. = free data point used to estimate the parameter)
  ⤳ 1 data point is sacrificed every time we estimate a parameter
- using R^2 would not punish "having too many regressors."
- we use adjusted - R^2 of R^{-2} when we want to punish adding too many regressors

$$R^2 = 1 - \frac{SSR}{SST} = \frac{1 - SSR/\cancel{n}}{SST/\cancel{n}}$$

$$\text{adj. } R^2 = \frac{1 - SSR/(n-k-1)}{SST/(n-1)}$$

If we have more k, d.f. = n-k-1 ↓, SSR /(n-k-1) ↑, adj - R^2 ↓

Using adjusted R-squared to choose between non-nested models (one model is not a subset of another).

Consider Model 1

$$\begin{array}{rcl} \widehat{salary} & = & \begin{array}{ccc} 830.63 & +0.0163 sales & +19.63 roe \\ (223.90) & (0.0089) & (11.08) \end{array} \\ n & = & 209, \quad R^2 = 0.029, \quad \bar{R}^2 = 0.020 \end{array}$$

Consider Model 2

$$\begin{array}{rcl} \log(\widehat{salary}) & = & \begin{array}{ccc} 4.36 & +0.2751 \log(sales) & +0.0179 roe \\ (0.29) & (0.033) & (0.004) \end{array} \\ n & = & 209, \quad R^2 = 0.282, \quad \bar{R}^2 = 0.275 \end{array}$$

← 27.5% of variation in Y is explained. So, this model is better!